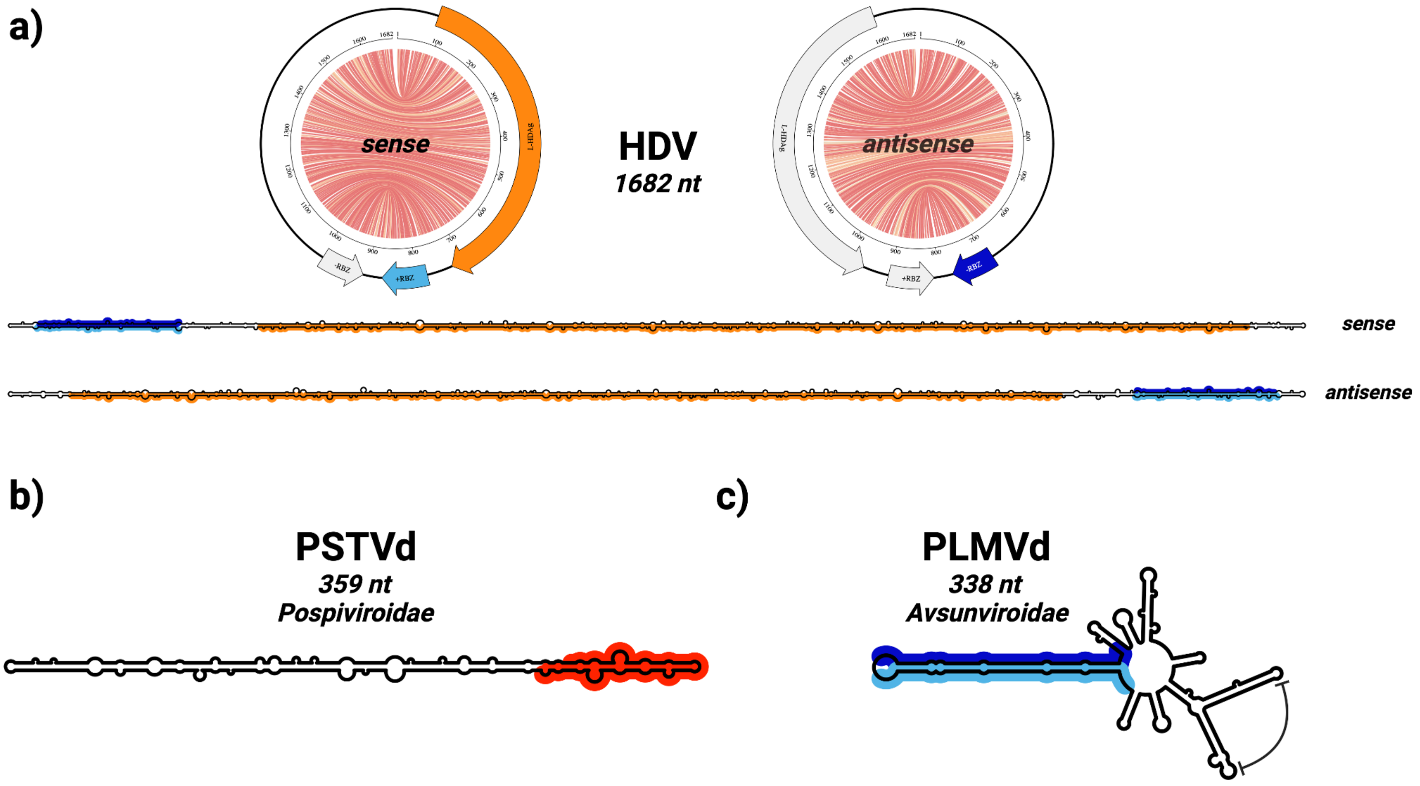
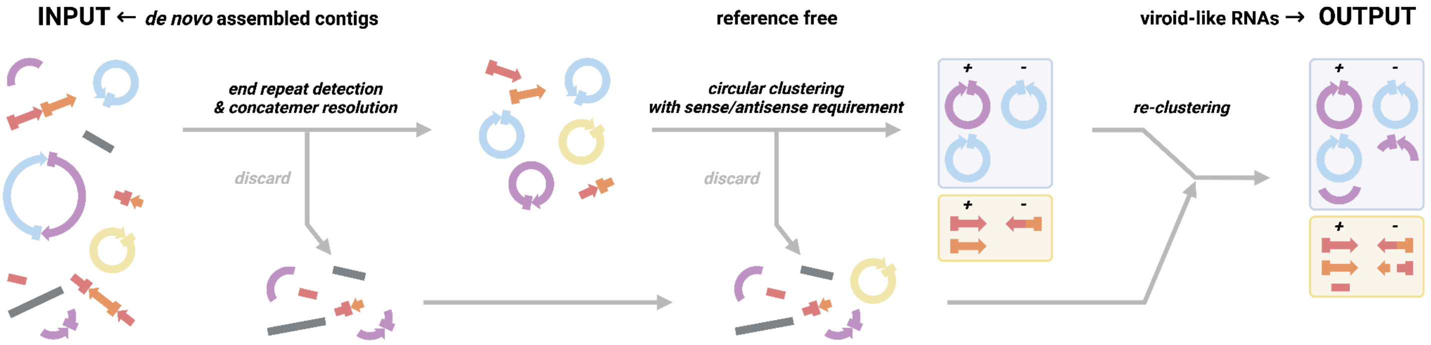
Supplementary Figures



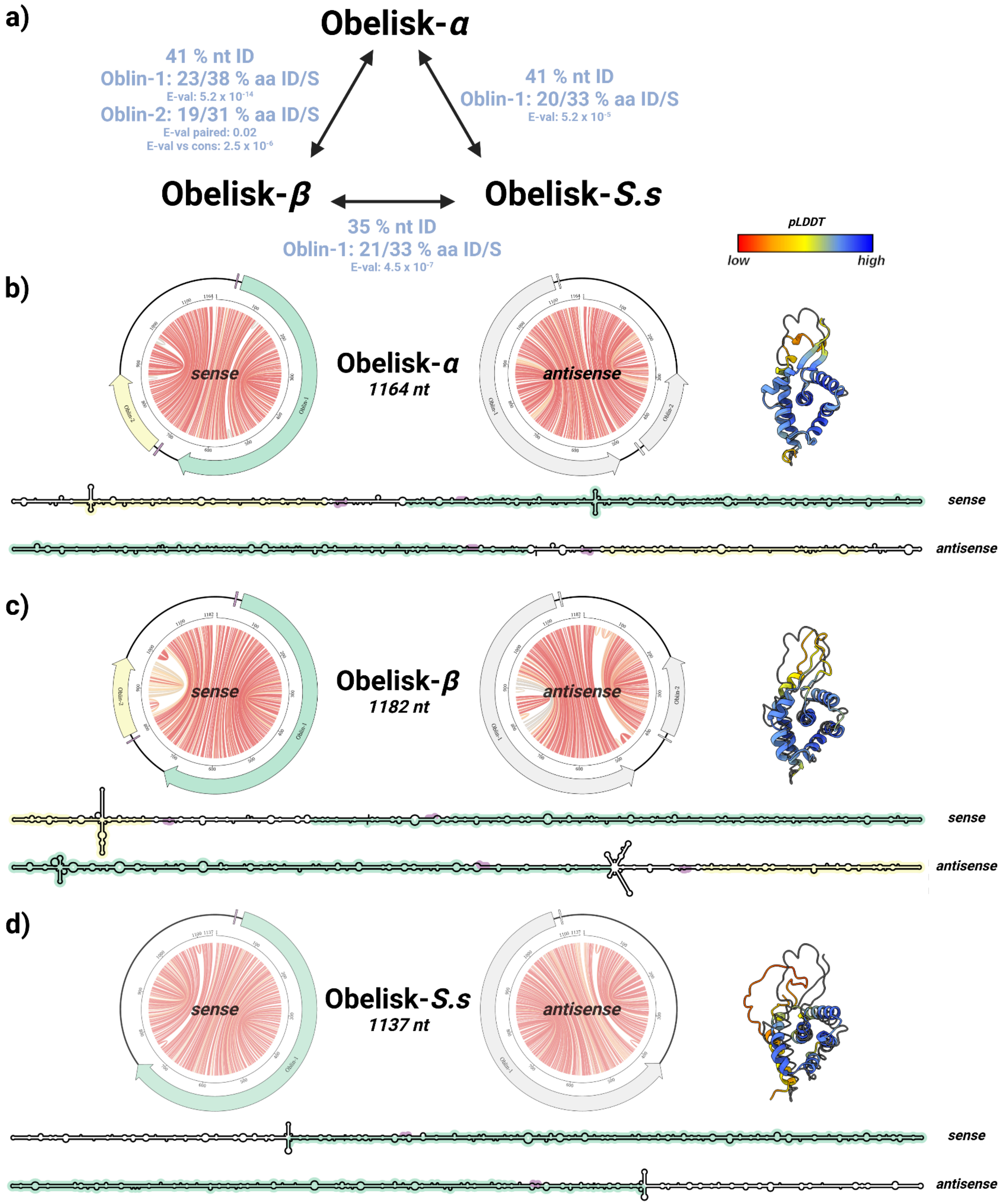
Supplementary Figure 1. Background on viroid and HDV families: Hepatitis delta virus, *Pospiviroidae*, and *Avsunviroidae* form a class of highly structured, circular sub-viral RNAs

a) the Hepatitis delta virus (HDV) genome (`NC_001653.2`) [114] is predicted to fold into a rod-shaped RNA secondary structure in both sense, and antisense - depicted here as both “jupiter” plots where chords represent predicted basepairs (coloured by basepair probability from 0, grey, to 1, red) with features greyed out in antisense, and “skeleton” diagrams. Large hepatitis delta antigen (L-HDAg, orange), and hepatitis delta ribozymes (RBZ, Rfam: `RF00094`, antisense: dark blue, sense: light blue) indicated. **b**) Potato spindle tuber viroid (PSTVd) of the family *Pospiviroidae* folds [115] into a rod-like RNA secondary structure similar to HDV but encodes no ORFs, though does possess a conserved Pospiviroid RY motif (Rfam: `RF00362`, red). **c**) Peach latent mosaic viroid (PLMVd) folds [116] into a highly basepaired, but “branched” RNA secondary structure as is characteristic of the *Avsunviroidae* family. Type III hammerhead ribozymes (Rfam: RF00008, antisense: dark blue, sense: light blue) and “P8” pseudoknot (curved flat-headed arrow) illustrated.


INPUT ← *de novo* assembled contigs
reference free
viroid-like RNAs → OUTPUT
end repeat detection & concatemer resolution
discard
circular clustering with sense/antisense requirement
discard
+
-
re-clustering
+
-
+
-
+
-

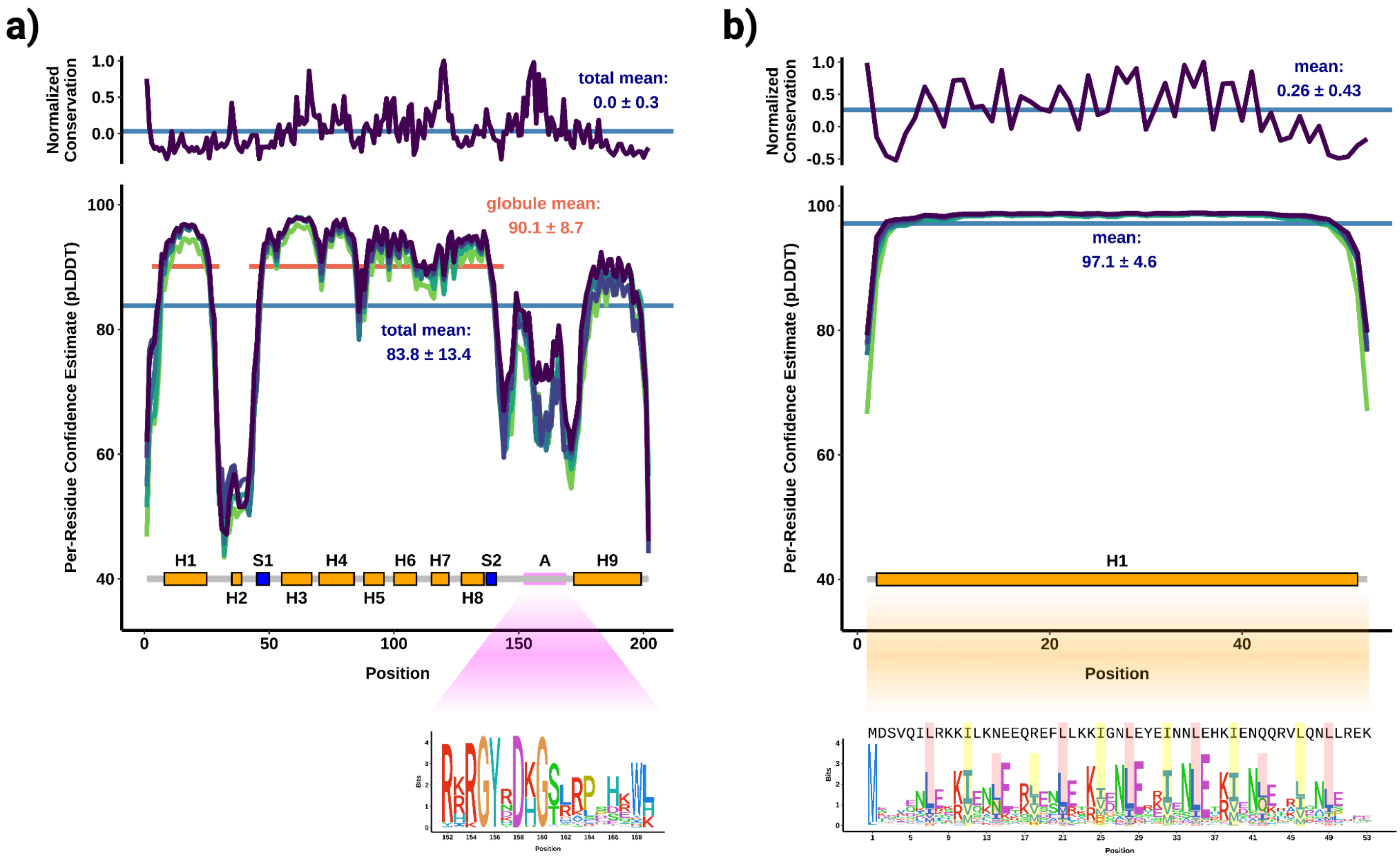
Supplementary Figure 2. VNom sequentially filters contigs to enrich for RNAs with viroid-like properties

VNom (short for "Viroid Nominator", pronounced *venom*) attempts to enrich for RNAs that are apparently circular and are present in the dataset in both polarities (a hallmark of RNA replication). To do this, VNom takes in *de novo* De Bruijn graph assembled contigs (from stranded RNA-seq data) and filters for potentially circular contigs by looking for perfect k-mer matches at the ends of each contig. Further, VNom also attempts to resolve concatemeric contigs by looking for regular repetition of such identified k-mers. These potentially circular contigs are then clustered based on sequence similarity using a circularly-permuting clustering algorithm. These resulting clusters are then kept if at least one contig of each polarity is identified by k-mer counting. Finally, these filtered clusters are compared against all of the previously discarded contigs to identify any remaining cluster members. While these filters should enrich for viroid-like RNAs, highly repetitive sequences also satisfy these requirements and so are often also enriched. VNom was found to work adequately well on deeply sequenced viroid-positive plant RNA-seq datasets (*e.g.* SRR11060618, SRR11060619, SRR11060620, and SRR16133646), especially when assemblies from the same bioProject were grouped together.


a)
Obelisk-α
41 % nt ID
Oblin-1: 23/38 % aa ID/S
E-val: 5.2 x 10-14
Oblin-2: 19/31 % aa ID/S
E-val paired: 0.02
E-val vs cons: 2.5 x 10-6
41 % nt ID
Oblin-1: 20/33 % aa ID/S
E-val: 5.2 x 10-5
Obelisk-β
Obelisk-S.s
35 % nt ID
Oblin-1: 21/33 % aa ID/S
E-val: 4.5 x 10-7
pLDDT
low
high
b)
sense
Oblin-1
Oblin-2
Obelisk-α
1164 nt
antisense
sense
antisense
c)
sense
Oblin-1
Oblin-2
Obelisk-β
1182 nt
antisense
sense
antisense
d)
sense
Oblin-1
Obelisk-S.s
1137 nt
antisense
sense
antisense

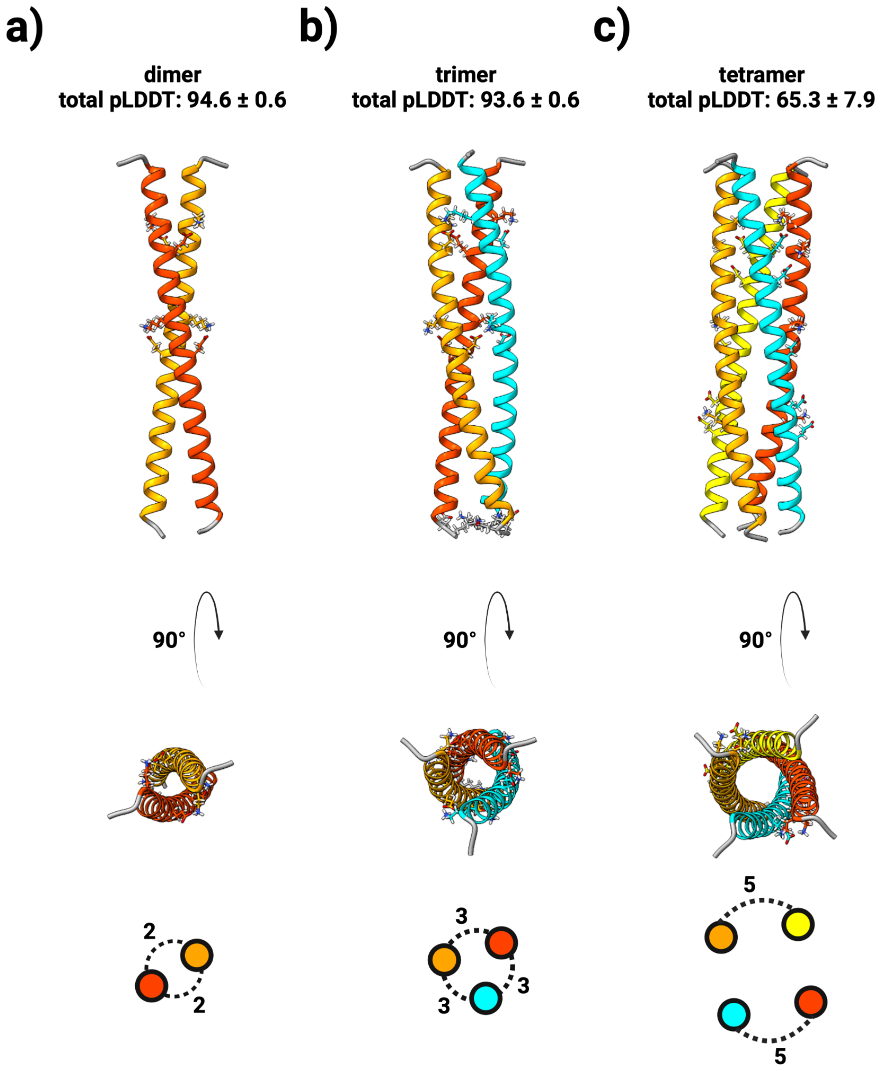
Supplementary Figure 3. Obelisks -*alpha*, -*beta*, and -*S. sanguinis* appear to belong to the same, diverse family

a) nucleotide (nt) and amino acid (aa) -level pairwise sequence identities (ID) and similarities (S) between Obelisks- α, β, and *S.s*. For Oblin protein sequences, mean pairwise `blastp` E-values are shown. Note, for Oblin-2 the pairwise `BLASTp` E-value relative to the Oblin-2 consensus (see methods) is also shown, indicating a distant, but evident homology between the α and β Oblin-2s. **b-d**) These Obelisks are similar in lengths; 1164, 1182, and 1137 nt, respectively, and share globally similar obelisk-like predicted RNA secondary structures in both their sense and antisense - depicted here as both “jupiter” plots where chords represent predicted basepairs (coloured by basepair probability from 0, grey, to 1, red) with features greyed out in antisense, and “skeleton” diagrams. Likewise, the genomic synteny of predicted open reading frames (ORFs, preceded by predicted Shine-Delgarno sequences, purple) appear to be shared, with Oblin-1 (green) consistently being present on one half of the predicted RNA secondary structure, and Oblin-2 (yellow), when present, following shortly after Oblin-1. `ColabFold` predictions of Oblin-1 tertiary “globule” structures built with *ad hoc* multiple sequence alignment (MSA) construction (coloured cartoons) superimposed over the `RDVA`-derived MSA prediction for Obelisk-α (black line, Figure 2a, see methods) indicating a conserved tertiary structure. Prediction confidence (pLDDT) shown as a colour bar (low confidence: 0, red; high confidence: 100, blue).



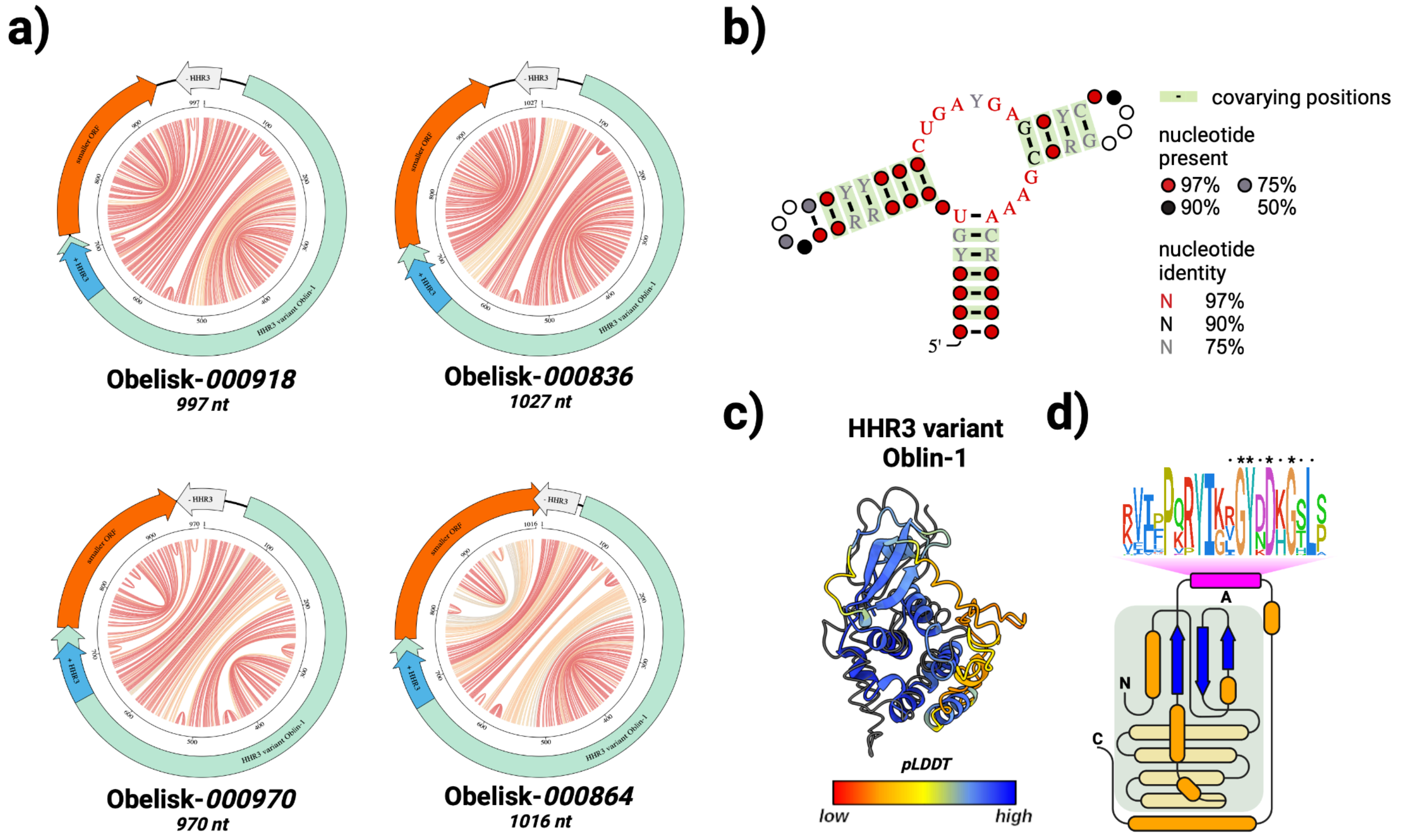
Supplementary Figure 4. Oblins are diverse and generate robust protein fold predictions

a) normalised conservation (top, above zero = more conserved, see methods) of Obelisk open reading frame 1 (Oblin-1) relative to Obelisk-ɑ indicates that Oblin-1 is largely poorly conserved (mean per-residue confidence estimate, μ-pLDDT ± standard deviation of 0.0 ± 0.3) but has three regions of conservation, around the C-termini of alpha helices 3 and 7, and *domain-A* (see sequence logo callout, bottom). Oblin-1 tertiary structure prediction per-residue confidence estimate (bottom, see methods) suggests a medium confidence total fold (μ-pLDDT: 83.8 ± 13.4), and a high confidence N-terminal "globule" (μ-pLDDT: 90.1 ± 8.7) that is consistently predicted over the top five models (green lines). *domain-A* is consistently predicted without a confident tertiary structure. **b**) Obelisk Oblin-2 has a higher mean normalised conservation (top, 0.26 ± 0.43), and is confidently predicted to form an alpha helix (μ-pLDDT: 97.1 ± 4.6). The Oblin-2 sequence logo (callout, bottom) shows leucine zipper features with "i+7" leucine spacing emphasised in red, with hydrophobic "d" position residues emphasised in yellow (Obelisk-ɑ Oblin-2 sequence shown for reference). Obelisk-ɑ alpha helices (orange boxes, "H" labels), and beta sheets (blue boxes, "S" labels) illustrated for clarity.



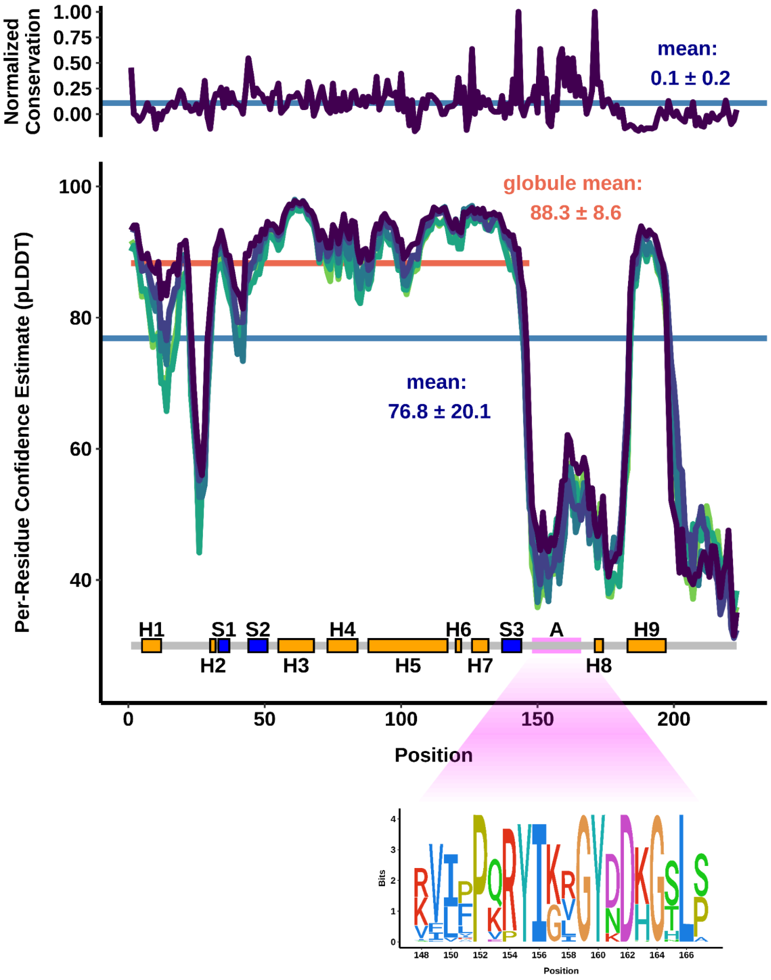
Supplementary Figure 5. Oblin-2 is predicted to homo-multimerize

tertiary structure predictions of Obelisk-*alpha* open reading frame 2 (Oblin-2) homo-multimers: **a**) dimer (mean pLDDT ± standard deviation: 94.6 ± 0.6), **b**) trimer (mean pLDDT: 93.6 ± 0.6), and **c**) tetramer (mean pLDDT: 65.3 ± 7.9). Residues involved in inter-helix salt bridges emphasised, and salt bridge counts illustrated on bottom.



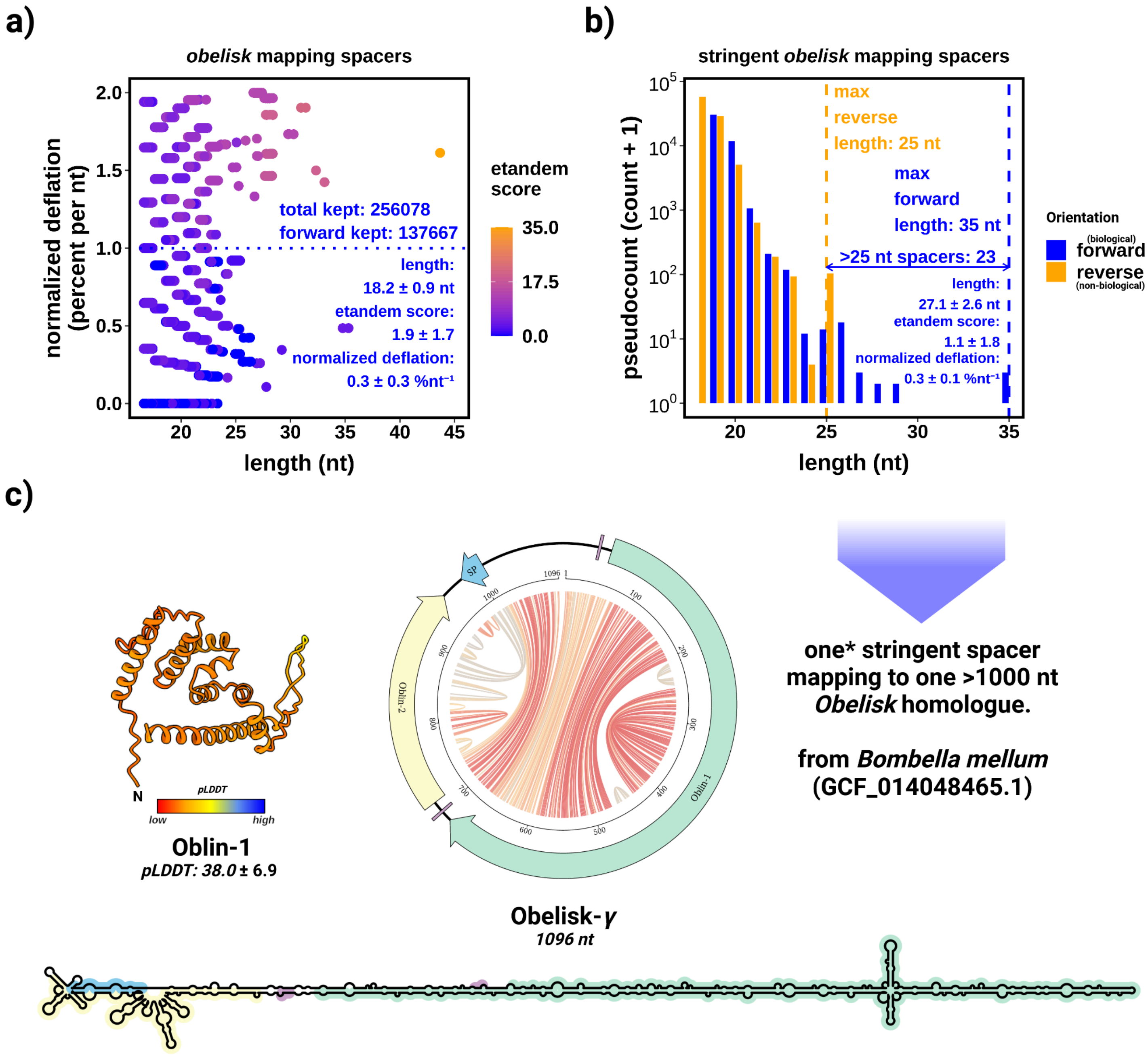
Supplementary Figure 6. Ribozyme-baring Obelisks encode a diverged Oblin-1

a) four “Obelisk-variant hammerhead type-III” (ObV-HHR3) -positive Obelisk genomes from Supplementary Table 1, illustrated as “jupiter” plots where chords represent predicted basepairs (coloured by basepair probability from 0, grey, to 1, red), Oblin-1 homologues illustrated in green, smaller, non-Oblin-2 ORFs in orange, and sense ObV-HHR3 in blue (with antisense ObV-HHR3 in grey). Note the conspicuous placement of ObV-HHR3 relative to Oblin-1 and the smaller ORF. **b**) the RDVA-derived, stringently-thresholded ObV-HHR3 covariance model summarised as a secondary structure with bairspair-forming, significantly covarying positions indicated with a green highlight. IUPAC “ambiguity codes” [117] used to represent RNA diversity: Y = U or C, R = A or G. **c**) ColabFold prediction of the “HHR-variant” Oblin-1 tertiary (“*Obelisk_000918*” as the reference sequence) structure built with a custom multiple sequence alignment (MSA) construction (coloured cartoons) superimposed over the RDVA-derived MSA prediction for Obelisk-α where possible (black line, Figure 2a, see methods). Prediction confidence (pLDDT) shown as cartoon colouring as in Supplementary Figure 3. **d**) a to-scale (secondary structure) topological representation of “HHR-variant” Oblin-1 with the “globule” shaded in grey (as in Figure 2b), and the *domain-A* emphasised with this bit-score sequence logo (see methods). Conserved “GYxDxG” motif emphasised.



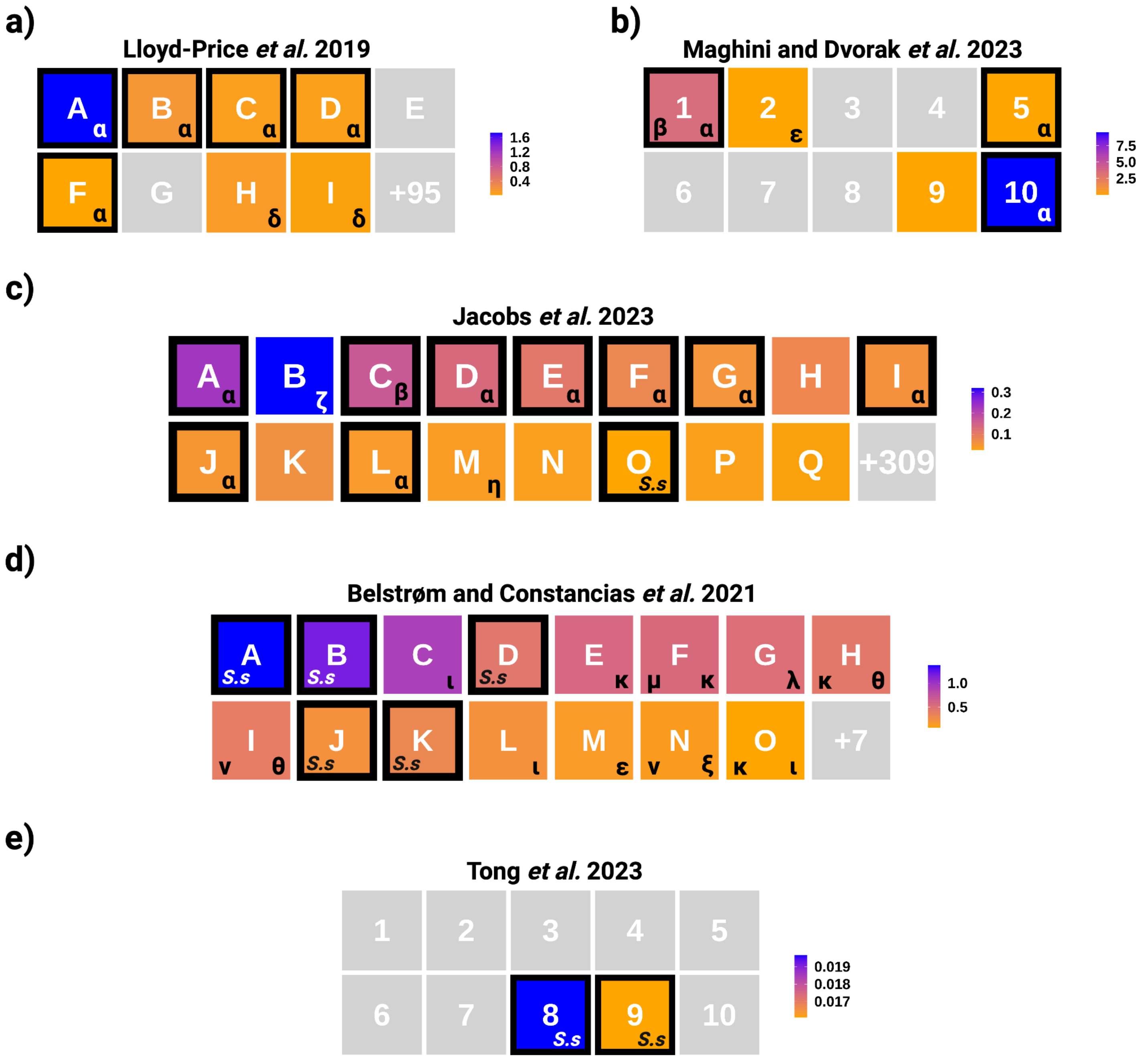
Supplementary Figure 7. Ribozyme-variant Oblin-1 has similar tertiary fold prediction characteristics to conventional Oblin-1s

normalised conservation (top, above zero = more conserved, see methods) of "Obelisk-variant hammerhead type-III" (ObV-HHR3) "HHR3-variant" Oblin-1 indicates that, similarly to the non-HHR3 Oblin-1 (Supplementary Figure 4), the "HHR3-variant" Oblin-1 is largely poorly conserved (mean normalised conservation ± standard deviation: 0.1 ± 0.2) but retains a conserved *domain-A* (see sequence logo callout, bottom). "HHR3-variant" Oblin-1 tertiary structure prediction per-residue confidence estimate (bottom, see methods) suggests a medium confidence total fold (mean per-residue confidence estimate, μ-pLDDT ± standard deviation of 76.8 ± 20.1), and a higher confidence N-terminal "globule" (μ-pLDDT: 88.3 ± 8.6) that is consistently predicted over the top five models (green lines). *domain-A* is consistently predicted without a confident tertiary structure.



Supplementary Figure 8. No evidence for capture of Obelisk sequences in available CRISPR-array data

a) an x-axis "jittered" scatter plot of Obelisk k-mers that map to the IMG/M spacer database [36] arranged by a proxy of information content (length-normalised percent deflation, lower = less deflated = more information), coloured by a metric of internal k-mer repetitiveness (see methods). Mappings with a length normalized deflation less than 1.0 percent per nucleotide were kept. Both mappings to "forward" and "reversed" (*not* reverse complemented) Obelisks were kept. Summary statistics on kept k-mers shown in bottom right hand corner. **b**) bar chart representing the noise floor to k-mers kept from a). 23 "forward" mapping k-mers (blue) longer than the longest "reverse" mapping k-mers (orange, 25 nt) were kept. Mappings below this threshold cannot be distinguished from noise. Summary statistics for these kept "forward" k-mers shown in the bottom right hand corner. **c**) ultimately one >1000 nt Obelisk genome was retrieved with two k-mer mappings to the same spacer locus (so the same spacer, see methods). This 1096 nt Obelisk-“gamma” (Obelisk-ɣ) exhibits a “rod-like” predicted secondary structure (“jupiter” plot, centre, “skeleton” diagram, bottom) and contains homologues to Oblin-1 (green) and Oblin-2 (yellow), with the spacer mapping to position ~1000 (steel-blue “SP” on the jupiter plot). The Obelisk-ɣ Oblin-1 is not predicted to fold into the characteristic “globule” tertiary structure (Figure 4 - tertiary structures). The “frayed” end where the spacer maps deviates from the “rod-ness” of other Obelisks (Figure 4 - “jupiter” plots), suggesting that this Obelisk-ɣ genome might be a chimeric mis-assembly.



Supplementary Figure 9. Human gut and oral microbiomes harbour diverse Obelisks

Heatmaps of Obelisk positive donors (>10 reads, averaged over donor if multiple samples) as inferred by k-mer and Oblin-1 pHMM matching (see methods and Table 5, donors with complex internal nomenclature were re-named for clarity see Table 6). Samples emphasised with black boxes were k-mer positive (but not exclusively). Lowercase Greek lettering indicate which Obelisks were found in a given donor as inferred by either k-mer counting (black boxes - k-mer profiling Obelisks -α, -β, and -*S.s*), or by *post hoc* classification of newly assembled and independently clustering Obelisks (see methods). Human gut microbiome samples: **a**) *Lloyd-Price et al. 2019* [20], **b**) *Maghini and Dvorak et al. 2023* [79], and **c**) *Jacobs et al. 2023* [112]. Human oral microbiome samples: **d**) *Belstrøm and Constancias et al. 2021* [37], and **e**) *Tong et al. 2023* [113]. Colour scales indicate Obelisk read counts relative to total donor reads x10^{-4}. Greek letter key: α : alpha, β : beta, δ : delta, ε :

epsilon, ζ : zeta, η : eta, θ : theta, ι : iota, κ : kappa, λ : lambda, μ : mu, ν : nu, and ξ : xi. Obelisks diagrammed in Figure 4.

Supplementary Table 1. see Data Availability

A unified set of Obelisk RNAs grouped hierarchically by percent identity (`circUCLUST` default settings). To ensure stringency, only full length genomes from the `RDVA` dataset were used (subset at 700 nt ≤ length ≤ 2000 nt), as identified by `CircleFinder` (`VNom` settings). Genomes were clustered first at the 80 % identity level, which we define as the boundary between Greek lettering, then at the 95 % identity level, which we define as the sub-type threshold. Open reading frames were then predicted (`prodigal`, `-p meta`) and genomes were converted to match the strand polarity of the largest predicted ORF, placing the first nucleotide of the start codon at the 51st nucleotide. 1,744 80 % identity stringent clusters (composed of 7,202 genomes total) were found. A naming convention is proposed with the following pattern *"Obelisk_X_Y_Z"* where "X" refers to the 80 % cluster ordinate, "Y" to the 95 % cluster ordinate, and "Z" as a unique identifier within the 95 % cluster. The first 15 80 % ordinates are defined as the Obelisks depicted in Figure 4, the next 10 80 % ordinates are defined as the remaining letters in the Greek alphabet (*omicron* through *omega*). As such, the centroid Obelisk-ɑ sequence that is also the centroid of the first 95 % sub-type is defined as "*Obelisk_000001_000001_000001"*. For completeness, an equivalent, additional clustering (see Data Availability) of the `RDVA` dataset without the `CircleFinder`, or `prodigal` steps (subset at 700 nt ≤ length ≤ 1500 nt) is provided. This clustering yielded 6108 80 % clusters of 14,235 genomes total. We caution that this dataset is more likely to be mis-clustered due to unaccounted-for peculiarities of *de novo* assembly, and issues arising from clustering arbitrary reverse-complemented sequences, as such, please use the clusterings (and numberings) in Supplementary Table 1 as the starting point for further Obelisk characterization.

Supplementary Table 2. see Data Availability

The *domain-A* alignment and metadata used to construct, and annotate Figure 3.